Clasificación por Enterotipos y Grupos Ortólogos del Microbioma Humano con Métodos No Supervisados

Cristóbal R. Santa María*, Victoria Santa María**, Laura Ávila*,
Juan Otaegui*, Marcelo Soria***
*DIIT-UNLaM, **Instituto Lanari-FMed-UBA, ***FAUBA
Florencio Varela 1903 San Justo Pcia. de Buenos Aires 54-011-44808952
csantamaria@unlam.edu.ar vctrsntmr@hotmail.com
laura_avila75@yahoo.com.ar soria@agro.uba.ar
juancarlosotaegui@yahoo.com.ar

Resumen

Se relatan las tareas llevadas a cabo por el Grupo de Investigación y Desarrollo en Data Mining del Departamento de Ingeniería e Investigaciones Tecnológicas de la UNLAM durante el año 2015 en el marco del Proyecto de Incentivos C169 "Aplicaciones de Data Mining al Estudio del Microbioma Humano". Se detallan las pruebas realizadas con el software SUPERFOCUS y la base de datos genéticos SEED, para desarrollar los análisis taxonómicos y funcionales que prepararan la información de las secuencias microbiómicas para procesarla con algoritmos de data mining. Se explicitan los aspectos teóricos y prácticos de la aplicación de estos algoritmos sobre conjuntos de prueba. Se analiza la interpretación clínica dada a los resultados y finalmente se describen los cursos de acción para continuar con la investigación durante 2016.

Palabras Claves: ADN, Microbioma, Enterotipos, Grupo Ortólogo, Clasificación.

Contexto

En esta línea de investigación se intenta construir un algoritmo de clasificación de estadíos de desarrollo de cáncer de colon y enfermedad de Crohn basado en la información aportada por el ADN constituyente del microbioma humano, en particular el intestinal. A partir de la secuenciación del ADN microbiano presente en el intestino, cada secuencia genética es una instancia en una base de datos sobre la que es posible aplicar procedimientos de aprendizaje no supervisado para agrupar las secuencias correspondientes a un gen marcador por especies u otros taxones más generales. Sin embargo tales categorizaciones suelen tener un sesgo, respecto de la caracterización clínica, que es producto de la secuenciación misma y de las técnicas que se aplican previas al agrupamiento. Por tal motivo conviene explorar agrupamientos de todas las secuencias genéticas, y no ya solo las de un gen marcador, realizados también en forma no supervisada, de acuerdo a la función que les corresponda en el metabolismo y que resulta distinta en la salud y en cada estadío de la enfermedad. Se espera que el estudio realizado interrelacionando ambos tipos de agrupamientos proporcione categorías de clasificación estables y compatibles con las caracterizaciones clínicas de los estadíos de salud y enfermedad. Con tales categorías se intentará a continuación aplicar métodos de aprendizaje supervisado como ensambles de árboles de decisión para obtener un clasificador que colabore en la clínica de prevención, diagnosis o prognosis. Se pretende además elaborar como resultado de los pasos descriptos una "pipeline" para investigar la aplicación de técnicas de data mining al microbioma humano en el caso de una enfermedad en general.

Introducción

Con vistas a la línea de investigación que el presente proyecto abre y que incluye la posibilidad futura de usar muestras propias, se decidió adaptar un servidor con el sistema BioLinux que posee una fácil interacción con paquetes de software libre utilizados en la investigación en biología computacional. Ocasionalmente se usaron también máquinas virtuales armadas en la nube para llevar adelante los procesos.

Al tener presente el objetivo de desarrollar una "pipeline" que en primer término instrumente una clasificación de microbiomas ajustada según criterios clínicos y luego utilice las categorías obtenidas para predecir en forma de diagnóstico o pronóstico las enfermedades estudiadas, se decidió llevar a cabo un ensayo sobre los procedimientos más citados para estas acciones. Para cobrar idea del volumen de datos a utilizar se comenzó estableciendo que espacio promedio, calculado en bytes, ocupa una secuencia. Los relevamientos de ADN total de microbiomas producen fragmentos cortos de ADN. Dependiendo de la tecnología de secuenciación y la configuración de los equipos se puede controlar el tamaño y cantidad de estos fragmentos. Con los equipos MiSeq y HiSeq de Illumina, los más usados hoy en día, las secuencias de ADN pueden tener entre 150 y 300 bases de largo y la cantidad puede llegar a varios millones de secuencias por corrida. Es común en el análisis de microbiomas tener unos 30 millones de secuencias por cada muestra analizada. Esto determina que sean críticos los tiempos de ejecución de los programas utilizados para el análisis.

Existen diferentes tipos de análisis para estudios de secuenciación metagenómica, por ejemplo, obtener "contigs", que son ensambles de secuencias en un fragmento bastante mayor y que permite la reconstrucción de genes completos, genomas parciales o, incluso, genomas completos. Otras aproximaciones posibles son la asignación taxonómica de las secuencias, es decir, determinar con la mayor precisión posible a que especie, género, familia, etc. de microorganismos pertenecen las secuencias obtenidas; o determinar una asignación funcional, que consiste en encontrar para las secuencias que codifican proteínas, de qué tipo son y en qué actividades celulares participan.

Líneas de Investigación, Desarrollo e Innovación

La línea de trabajo elegida se enfoca en las asignaciones taxonómicas y funcionales [1]. Para esto se realizaron pruebas con un software de reciente desarrollo, SUPERFOCUS, que efectúa la determinación taxonómica y la asignación funcional. Para esta última actividad el software utiliza la base de datos SEED [2] que asigna una función a cada proteína. SUPERFOCUS corre en plataformas Linux y está diseñado para aprovechar las características multiprocesador de las máquinas modernas. Las pruebas iniciales se realizaron en el servidor BioLinux con ocho procesadores Intel-I7 y 16 GB de memoria RAM. Se continuaron las pruebas en el servicio de computación en la nube con una máquina virtual Linux. Se determinó que con archivos de 500,000 secuencias el proceso se completaba con éxito en todas las pruebas. Se diseñó entonces un script de comandos Unix que particiona el archivo a procesar en bloques de 500,000 secuencias cada uno y las envía a SUPERFOCUS. Se procesaron cinco muestras diferentes, cada una con unos 26 millones de secuencias.

Las salidas de SUPERFOCUS son varios archivos de texto, uno con las asignaciones funcionales, tres con la información de subsistemas y otro más con la información taxonómica.

El procedimiento experimentado permite entonces conocer la distribución de frecuencias de genes o funciones de cada

paciente con la cual se realizará la clasificación en clusters de sus respectivos microbiomas.

Resultados y Objetivos

En [3] se analiza la influencia de la dieta en el cáncer de colon mediada por la composición de la microbiota a partir de muestras de población afroamericana (AA) y africana nativa (NA). Se miden las diferencias de microbiota de los AA y los NA. La representación numérica de cada microbioma intestinal se realiza entonces por un vector donde cada componente expresa la abundancia de cada microorganismo en él.

En el caso de las muestras AA y NA se trata de dos matrices integradas por las filas que representan los microbiomas de los individuos integrantes. El número de componentes del vector-microbioma es de unos cientos. Para visualizar en el caso presentado las diferencias entre los microbiomas AA y NA se apela a una reducción de variables por componentes principales resultando la Figura 1. De acuerdo a lo expuesto hasta aquí el trabajo computacional consiste en obtener los datos secuenciados de una muestra integrada por varios microbiomas

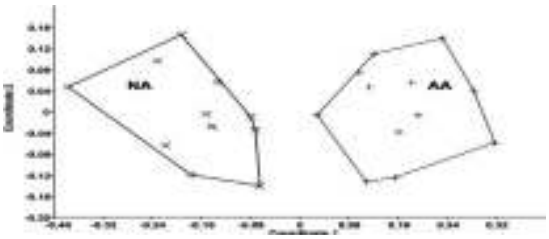

Figura 1

Cada microbioma de esta muestra debe cotejarse con una base de datos correspondiente a un gen marcador para encontrar la distribución de frecuencias de los microorganismos identificados por tal gen. Alternativamente el conjunto de secuencias del microbioma puede compararse con otra base de datos de funciones genéticas para agrupar los genes integrantes por función y así obtener la distribución de frecuencias según las funciones metabólicas que las secuencias integrantes revelan. [1]. En cualquier caso los microbiomas individuales pueden agruparse en clusters. Los conjuntos obtenidos clasifican a los individuos según características clínicas cuyo valor debe ser sopesado desde el punto de vista médico. Si la clasificación tiene significancia clínica puede utilizarse con métodos predictivos tales como árboles de decisión para establecer diagnósticos o pronósticos en pacientes aún no clasificados [4]

A partir de [3] y a efecto de estudiar la metodología empleada, se obtuvieron dos conjuntos de datos A y B que agrupan a los microbiomas intestinales de 33 pacientes. Cada fila del conjunto A representa el microbioma de un paciente obtenido en base al gen marcador. Para el conjunto B cada fila representa el microbioma de un paciente y cada columna una característica o grupo funcional hallada.

Se continuó realizando la clasificación en clusters. Para medir la distancia entre dos muestras, dos microbiomas cuyas distribuciones de especies o géneros son conocidas se consideró que cada paciente está representado por una distribución de frecuencias estadísticas. Hay que medir si las dos distribuciones se parecen a efecto de terminar ubicándolas en clusters. Para ello se utilizó la distancia de Jensen-Shanon [5]

$$D_{PQ} = \left(\sum_{i=1}^{N} p_i \log p_i + q_i \log q_i\right)^{\frac{1}{2}}$$

Esta distancia es la que usa el algoritmo PAM (Partitioning around medoids) utilizado para este caso directamente de la biblioteca cluster de R. El medoide es el elemento para el cual la disimilitud promedio con todos los objetos en el conglomerado es mínima. Se estudió también, con la idea de programarlo como alternativa más rápida el algoritmo

propuesto en [6] que trabaja con medoides pero en forma similar a k-means.

Para definir el número óptimo de clusters se calculó el índice de Calinski-Harabasz : $CH = \frac{\frac{B_k}{k-1}}{\frac{W_k}{n-k}}$

Aquí B_k es la suma de las distancias al cuadrado de todos los elementos *i* y *j* que no pertenecen al mismo cluster, W_k es la suma de los cuadrados de las distancias de todos los elementos *i* y *j* que pertenecen al mismo cluster, *n* es el número de elementos a clasificar y *k* la cantidad seleccionada de clusters. Utilizando el comando nclusters de la biblioteca clusterSim de R se pudo programar el testeo de CH para distintos valores de *k* a fin de hallar la cantidad óptima de clusters. En la Figura 2 se ve un ejemplo para el que se ha utilizado el conjunto A. El número de clusters $k = 3$ produce el mayor valor del índice CH.

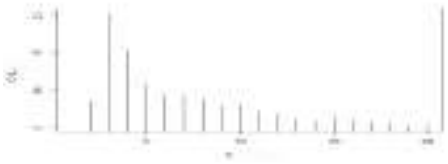

Figura 2

Con el número de clusters obtenido se corre el algoritmo PAM para obtener los enterotipos en cuestión. Desde el punto de vista computacional la consistencia de tal agrupamiento se mide con el índice Silhouette [7] Sin embargo luego de las pruebas de agrupamientos realizadas se requirió evaluar el aspecto clínico de los resultados obtenidos.

Hay más de 1000 especies de microbios que viven en el intestino humano y conforman el microbioma. Este juega un rol importante en la protección del huésped contra patógenos, modula la inmunidad, regula procesos metabólicos y es incluso considerado en algunos casos como un "órgano endocrino". Secuenciando 16s rADN ha sido sugerido que el microbioma intestinal puede dividirse en tres diferentes "enterotipos" [3], cada uno de ellos pueden ser identificados por la variación en los niveles de alguno de los siguientes tres géneros: Bacteroides (enterotipo 1) Prevotella (enterotipo 2) y Rominococcus (enterotipo 3). Sin embargo estos enterotipos no se definen tan claramente como por ejemplo los grupos sanguíneos, no parecen ser distintos con respecto a su riqueza funcional y no han podido ser correlacionados a características poblacionales tales como la edad, el IMC o la nacionalidad.

El estudio de biomarcadores funcionales arrojo mejores resultados, hallándose correlación entre algunos de ellos y propiedades del huésped lo que puede ser útil para el diagnóstico y prevención de enfermedades. El resultado final de este análisis se utiliza para describir un número potencial de funciones y su relativa abundancia en el metagenoma.

Para realizar estos estudios en vez de dividir al microbioma en diferentes enterotipos por clasificación taxonómica se lo divide en grupos ortólogos (OG) que codifican para distintas enzimas o proteínas pertenecientes a distintas vías metabólicas. [3] Para la Enfermedad de Crohn se ha encontrado que su manifestación está asociada a una disminución de las especies bacterianas (Bacteroides y Clostridium) en la mucosa colónica y aumento de otras tales como las fusobacterias. Sin embargo esto no se refleja en las muestras tomadas de la materia fecal [8]. La etiología del cáncer colorectal es multifactorial. Se ha encontrado una abundancia relativa de Bactteroidaceae, Streotococcaceae, Fusobacteriaceae, Peptostreptococcaceae, Veillonellaceae y Pasteurellaceae y una disminución en los niveles de Lachnospiraceae, Ruminococcaceae y Lactobacillaceae en los pacientes con

cáncer colorrectal [9]. Además el microbioma induce un estado de inflamación crónica y genera metabolitos reactivos y carcinógenos que contribuyen al desarrollo del cáncer colorrectal. Se ha trabajado sobre la hipótesis de que la utilización de biomarcadores del microbioma y el análisis de otros factores de riesgo pueden contribuir al screening de cáncer colorrectal. En este caso se encuentran diferencias significativas entre los sujetos normales y los distintos estadios de cáncer colorrectal [10]. También estas disbiosis favorecen la prevalencia de bacterias dentro del tumor que poseen genes altamente virulentos que contribuyen al desarrollo de la enfermedad [9].La división según características taxonómicas no parece ser la más adecuada dado que no se refiere a la funcionalidad de las bacterias y tiene una amplia variación inter-individuos, por lo que el clustering debe realizarse según grupos de co-abundancia que combinen a los distintos tipos de bacterias en función de las vías metabólicas asociadas a cada patología. Ya se ha analizado que existen vías metabólicas que resultan protectoras, otras que son desencadenantes de la enfermedad y otras que la perpetúan, probablemente esta sea la forma de empezar la clasificación. Por otro lado la flora de la materia fecal no es igual a la flora de la mucosa colónica por lo que el estudio sería más representativo si se tomaran muestras a partir de biopsia de mucosa colónica.

Formación de Recursos Humanos

En el grupo participan un Magister y un Especialista en Explotación de Datos, un Doctor en Biología, dos Médicos, dos Ingeniero en Sistemas, un Matemático y un alumno de la carrera de Informática. Actualmente hay en desarrollo una tesis de maestría.

Bibliografía

[1] Ngom-Bru, Catherine and Barretto, Caroline. Gut microbiota: methodological aspects to describe taxonomy and functionality. Briefings in Informatics. Vol3 NO 6. 747-750

[2] http://theseed.org

[3] Arumugam, M et al. Enterotypes of the human gut microbiome. Nature 2011 may 12; 473(7346): 174-180. doi:10.1038/nature09944

[4] Junhai et al. Diet, microbiota, and microbial metabolites in colon cancer risk in rural Africans and African Americans. Am. J. Clin. Nutr. 2013. 98. 11-120.

[5] Endres, D y Schindeling,J. A New Metric for Probability Distributions. IEEE. Transactions on Information Theory. Vol. 49 NO.7. 2003.

[6] Hae-Sang Park, Chi-Hyuck Jun. A simple and fast algorithm for K-medoids clustering. Expert Systems with Applications 36 (2009) 3336-3341.

[7] Rousseeuw, P. Silhouettes: a graphical aid to the interpretation ans validation of cluster analysis. Journal of Computational and Applied Mathematics 20 (1987) 53-65.

[8] Ray K. IBD. Understanding gut microbiota in new-onset Crohn's disease. Nat Rev Gastroenterol Hepatol [Internet]. Nature Publishing Group; 2014;11(5):268.

[9] Burns MB, Lynch J, Starr TK, Knights D, Blekhman R. Virulence genes are a signature of the microbiome in the colorectal tumor microenvironment. Genome Med [Internet]. 2015;7(1):55.

[10] Zackular JP, Rogers MAM, Ruffin MT, Schloss PD. The human gut microbiome as a screening tool for colorectal cancer. Cancer Prev Res (Phila) [Internet]. 2014;7(11):1112–21.